TouchSDF: A DeepSDF Approach for 3D Shape Reconstruction Using Vision-Based Tactile Sensing

Mauro Comi[1], **Yijiong Lin**[1,2], **Alex Church**[1,2], **Laurence Aitchison**[1], **Nathan F. Lepora**[1,2]
[1]University of Bristol, [2]Bristol Robotics Laboratory
mauro.comi@bristol.ac.uk

Abstract

Humans rely on their visual and tactile senses to develop a comprehensive 3D understanding of their physical environment. Recently, there has been a growing interest in manipulating objects using data-driven approaches that utilise high-resolution vision-based tactile sensors. However, 3D shape reconstruction using tactile sensing has lagged behind visual shape reconstruction because of limitations in existing techniques, including the inability to generalise over unseen shapes, absence of real-world testing and limited expressive capacity imposed by fixed topologies of graphs or meshes. To address these challenges, we propose TouchSDF, a Deep Learning approach for tactile 3D shape reconstruction that leverages the rich information provided by a vision-based tactile sensor and the expressivity of the implicit neural representation DeepSDF. This combination allows TouchSDF to reconstruct smooth and continuous 3D shapes from tactile inputs in simulation and real-world settings, opening up research avenues for robust 3D-aware representations and improved multimodal perception for robot manipulation. Code and supplementary material are available at: `https://touchsdf.github.io/`

1 Introduction

The current state of 3D shape reconstruction research is primarily concerned with the sense of vision [2] [17]. However, training Computer Vision algorithms for object manipulation is challenging due to the high-dimensional observation space, which suffers from occlusion, external lighting conditions and a distal view. Recently, several data-driven methodologies that utilise vision-based tactile sensors for 3D understanding have been proposed [5, 14, 8, 9]. To the best of our knowledge, Smith et al. [15] proposed the first and only approach for vision-based tactile reconstruction, which utilised DIGIT vision-based tactile sensors [7] to extract contact-rich information. Although that work serves as a foundation for tactile-driven 3D reconstruction, a lack of testing in a real-world scenario makes it challenging to assess its applicability to robotic manipulation.

In this work, we propose TouchSDF, a novel approach for 3D shape reconstruction that leverages implicit neural representations for vision-based tactile sensing. Unlike existing methods that utilise discrete representations, TouchSDF employs DeepSDF [13], an implicit neural representation that encodes a smooth and continuous surface, enabling more accurate and robust reconstructions. DeepSDF is capable of reconstructing 3D geometry from partial point clouds, which is useful in our context as tactile sensors provide partial observations of an object. A key step for 3D reconstruction in the real world is to have effective real-to-sim tactile image translation, where we extend previous work with a GAN-based method [3] to handle complex surfaces and 6D poses. Through evaluation on both simulated and real objects, we demonstrate the effectiveness of TouchSDF in tactile-based 3D shape reconstruction and highlight its potential for enhancing object exploration and manipulation tasks.

Contribution In summary, our work makes the following principal contributions:
1) We propose TouchSDF, an approach for 3D shape reconstruction using vision-based tactile

NeurIPS 2023 Workshop on Touch Processing: a new Sensing Modality for AI .

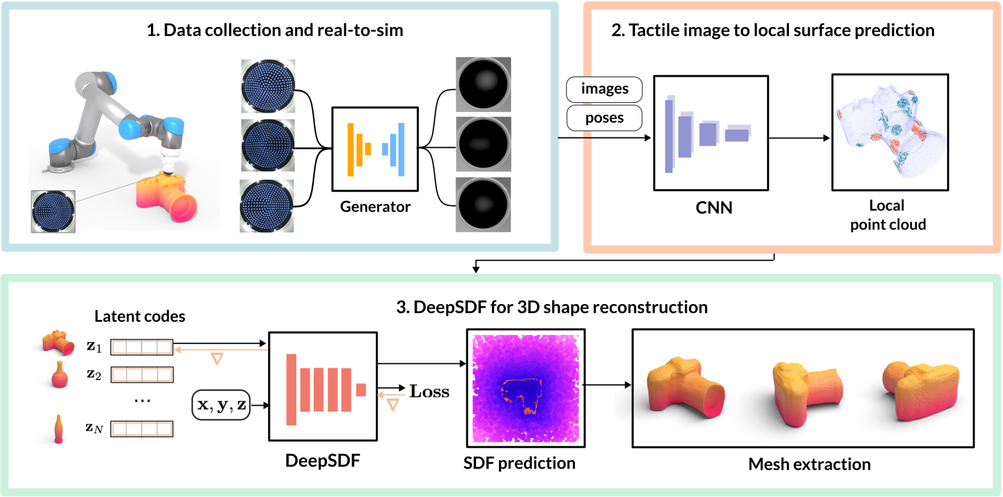
1. Data collection and real-to-sim
Generator
2. Tactile image to local surface prediction
images
poses
CNN
Local
point cloud
3. DeepSDF for 3D shape reconstruction
Latent codes
$\mathbf{z}_1$
$\mathbf{z}_2$
...
$\mathbf{z}_N$
$\mathbf{x}, \mathbf{y}, \mathbf{z}$
DeepSDF
Loss
SDF prediction
Mesh extraction

Figure 1: Overview of TouchSDF: (1) A robot samples the object's surface to obtain real tactile images that are translated into simulated images. (2) A Convolutional Neural Network (CNN) maps the simulated images to sets of 3D points representing the local object surface at the touch locations. (3) A pre-trained DeepSDF model predicts a continuous signed-distance function (SDF) representing the object shape from the point clouds over multiple contacts.

sensing. By leveraging the implicit neural representation DeepSDF, our method encodes smooth and continuous surfaces that enable accurate and robust reconstructions from partial observations.
2) We demonstrate the ability to generalise over unseen objects and poses both in simulation and in reality by conditioning on latent variables, thus encoding a wide range of geometries.
3) We give the first evaluation of 3D shape reconstruction using purely vision-based tactile sensing in a real-world setting.

2 Methodology

Our reconstruction procedure has three steps (Fig. 1). In the first step, we collect tactile images by sampling the surface of the object we intend to reconstruct. We also store the sensor (end-effector) pose at contact alongside the local point cloud of the surface at the touch location to serve as ground truth to train a local surface-prediction model. Real-world tactile images are converted into simulated tactile images using a real-to-sim GAN model [3]. In the second step, we predict a local point cloud based on the simulated tactile image and sensor pose. Finally, we employ a pre-trained DeepSDF model to reconstruct objects using the predicted partial point clouds.

2.1 Simulation

Collection of tactile images and local point cloud. Tactile images, local point clouds and sensor poses were collected for training purposes using a PyBullet-based tactile simulator (Tactile Gym [3] [10]). Specifically, the tactile readings collected consist of 256×256 pixel tactile images obtained by rendering the contact depth at the touch location using the PyBullet's synthetic camera. Collected local point clouds serve as the ground truth needed to train a model that maps a tactile image into the corresponding touched surface.

Tactile images to point cloud prediction. We applied Smith et al.'s approach [15] to map tactile images to local point clouds, using a simulated 6-DoF robot arm with a TacTip tactile sensor [3, 10]. Specifically, we collected 2D tactile images in simulation, representing depth maps of touched areas on objects, and their corresponding ground truth point clouds. We utilised a Convolutional Neural Network (CNN) to deform a base mesh into a predicted contact geometry, following the procedure outlined in[15]. Specificaly, we defined an initial base mesh, and then predicted vertex displacement using a CNN conditioned on tactile images. We then sampled a point cloud on the deformed mesh and optionally enhanced it by estimating surface normals and sampling additional points. This step increases the robustness of the DeepSDF model at inference time, as we do not only have points on the surface but also points inside and outside the shape. The CNN was trained to minimise the Chamfer Distance between the predicted point cloud and the ground truth point cloud.

3D shape reconstruction using DeepSDF. Following the methodology outlined in [13] on DeepSDFs, we constructed a dataset of 35000 SDF pairs $\{(\mathbf{x_j}, s_j)\}_{j=1}^{N}$ per object, where $\mathbf{x} \in \mathbb{R}^3$ are the coordinates sampled within a closed volume around the objects and $s_j \in \mathbb{R}$ corresponds to the associated signed distances. We employed the architecture originally proposed by [13] with added positional encoding [12, 18]. During training, the DeepSDF model learns a continuous signed distance function $f_\theta(\mathbf{x}, \mathbf{z})$ conditioned on a shape embedding $\mathbf{z}$. To ensure the encoding of shape information, an embedding $\mathbf{z_i}$ is jointly optimized with the network parameters θ for each shape i. As a result, similar shapes are encoded by similar latent vectors, enabling interpolations in the latent space that facilitate the reconstruction of unseen objects. The objective minimised during the training process follows [13]: $\mathcal{L}(\theta, \mathbf{z}) = \underset{\theta,\mathbf{z_i}}{\arg\min} \sum_i \sum_{j=1}^{N} ||f_\theta(\gamma(\mathbf{x_j}), \mathbf{z_i}) - s_j||_1 + \alpha \cdot ||\mathbf{z_i}||_2^2$. Here the regularisation term $||\mathbf{z_i}||_2^2$ weighted by α is crucial to avoid exploding gradients when optimising latent vectors.

Reconstruction. The CNN-predicted point cloud $\mathcal{O}$ provides a partial observation of the geometry at the touch location. Because these points lie on the object's surface, we attribute them with a signed distance value of zero. To reconstruct the complete shape of the target object, a DeepSDF auto-decoder [13] is conditioned on $\mathcal{O}$ to optimise a 128-dimensional latent code $\mathbf{z_i}$ by solving $\arg\max_{\mathbf{z_i}} \mathbb{P}(\mathbf{z_i}|\mathcal{O})$ via first-order optimisation. This is achieved by freezing the model's parameters θ and solving $\arg\min_{\mathbf{z_i}} \mathcal{L}(\mathbf{z_i})$. The inferred *global* latent vector represents a compact representation of the entire shape that best describes the partial observation. The signed-distance function is defined by conditioning a pre-trained DeepSDF model on coordinates $\mathbf{x}$ and inferred latent vector $\mathbf{z_i}$. Finally, a surface $\mathcal{S}$ is extracted by employing the deterministic Marching Cubes algorithm [11], which extracts the zero-level set of the predicted signed-distance function $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 | f_\theta(\mathbf{x}, \mathbf{z}) = 0\}$.

2.2 Real world

Hardware. We conducted real-world experiments with a 6-DoF industrial robot arm (ABB IRB 120). The robot was equipped with a high-resolution vision-based tactile sensor, for which we used a TacTip 3D-printed soft biomimetic tactile sensor [9]. For real-world evaluation, we 3D-printed four objects (two different bottles, a camera, and a bowl) selected from ShapeNetCore.V2 [1] and selected two everyday objects (a transparent jar and a mug). These objects are unseen during training.

Real-to-Sim Image Transfer. When reconstructing objects in the real world, the robot collects real tactile images, which the point cloud prediction model is not trained to process. Therefore, we map the real tactile images to simulated ones using the translation approach proposed in [3], which utilizes a Generative Adversarial Network (GAN) framework for image-to-image translation. The pix2pix GAN [6] is trained with pairwise simulated (depth map) and real tactile images (marker patterns) collected with the same contact poses [3] [10].

Sim-to-Real Object Reconstruction. In this step, we combine the point cloud prediction model, DeepSDF and real-to-sim image transfer to achieve sim-to-real object reconstruction. Firstly, we collect real tactile images by performing random contacts with the robotic arm-mounted tactile sensor onto the real-world objects. For each touch, the real image is translated into a simulated tactile image and labeled with the corresponding end effector pose. The CNN is used to predict a local point cloud describing the contact geometry. Finally, our pre-trained DeepSDF model is conditioned on the predicted point clouds and estimated signed distance to extract the shape of an object (see Sec. 2.1). The evaluation of our Sim-to-Real method is outlined in the Appendix.

3 Results

3.1 Simulation

Following [15], our evaluation was performed on the ABC dataset. We sampled 3500 shapes for training, 350 for validation, and 200 for testing. To ensure a fair comparison with Smith et al. [15], shapes were randomly sampled from the training, validation, and testing sets used by the authors. Results on an additional dataset (ShapeNet) are reported in the Appendix I. Table 1 shows the Chamfer Distance (CD) [16], Earth Mover's Distance (EMD) [4], and Surface error obtained by TouchSDF

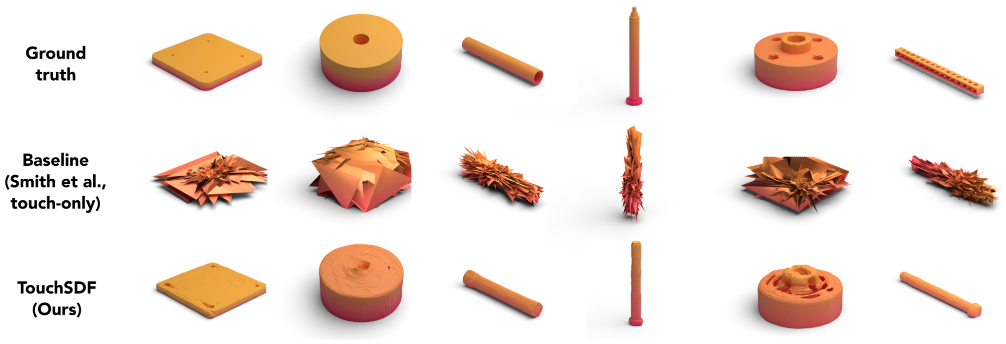

Figure 2: Qualitative comparison of TouchSDF and Smith et al. (grasp-only) performance after 20 touches.

and Smith et al. after 20 touches. To evaluate Smith et al.'s approach, we used the *grasp-only* model provided by the authors. Both EMD and Chamfer Distance were computed using 4096 points. TouchSDF achieves better EMD and surface reconstruction error, while achieving slightly lower CD despite a better visual quality (Fig. 2). This is due to has known limitations in accurately measuring the visual quality of reconstructed meshes [57] [58].

3.2 Real world

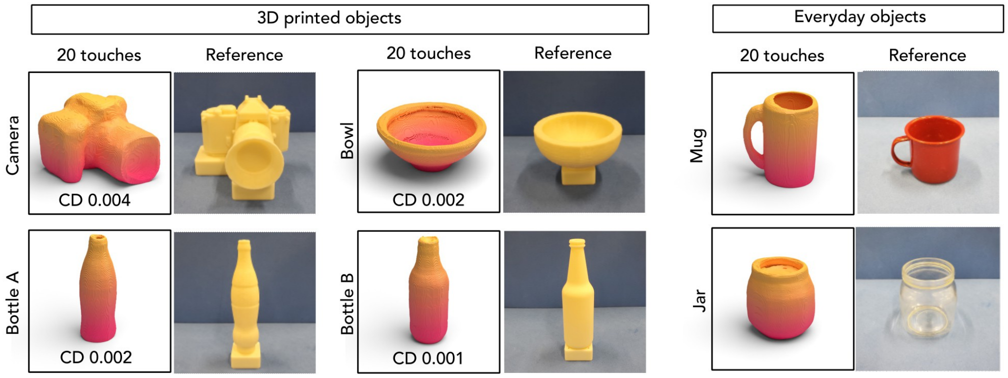

Figure 3: Reconstruction of four real 3D-printed objects and two everyday objects. The Chamfer Distance (CD) calculation requires a CAD model, hence it is applicable to 3D-printed objects but not to everyday objects. The cube-shaped mounting sockets are not part of the objects.

Reconstruction. To reconstruct the 3D-printed objects, we collected real tactile images from random locations on the object surfaces and translated them into their corresponding simulated images. The reconstruction procedure followed the same methodology employed in simulation. As the embeddings optimised by DeepSDF are not $SE(3)$-invariant, it was necessary to ensure that the pose of the object being reconstructed is consistent with the poses observed during training. To achieve this, the world reference frame was set manually to the centre of the objects. Our method successfully reconstructed both real 3D-printed objects, achieving a low Chamfer Distance that is comparable to those obtained in simulation (Fig. 3), and additional everyday objects (a mug and a transparent jar), for which the CD could not be computed due to the lack of CAD model. A further analysis on this result, as well as an ablation on the robustness to pose perturbations, is reported in the Appendix.

	CD (↓)	EMD (↓)	Surf. err. (↓)
Smith et al. (grasp-only)	**0.003**	0.19	2785%
TouchSDF (ours)	0.006	**0.07**	**36%**

Table 1: Comparison between TouchSDF and Smith et al.

Acknowledgements

This work was supported by an EPSRC PhD studentship on the ‘Interactive AI’ CDT program, a CSC-UOB joint PhD Scholarship, and a Leadership Award from the Leverhulme Trust on ‘A biomimetic forebrain for robot touch’ (RL-2016-39).

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[2] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer.

[3] Alex Church, John Lloyd, Nathan F Lepora, and others. Tactile sim-to-real policy transfer via real-to-sim image translation. In *Conference on Robot Learning*, pages 1645–1654. PMLR.

[4] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[5] Carolina Higuera, Siyuan Dong, Byron Boots, and Mustafa Mukadam. Neural contact fields: Tracking extrinsic contact with tactile sensing. *arXiv preprint arXiv:2210.09297*, 2022.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[7] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

[8] Mike Lambeta, Huazhe Xu, Jingwei Xu, Po-Wei Chou, Shaoxiong Wang, Trevor Darrell, and Roberto Calandra. Pytouch: A machine learning library for touch processing. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13208–13214. IEEE, 2021.

[9] Nathan F Lepora. Soft biomimetic optical tactile sensing with the tactip: A review. *IEEE Sensors Journal*, 21(19):21131–21143, 2021.

[10] Yijiong Lin, John Lloyd, Alex Church, and Nathan F Lepora. Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch. *IEEE Robotics and Automation Letters*, 7(4):10754–10761, 2022.

[11] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[12] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[13] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[14] Lukas Rustler, Jens Lundell, Jan Kristof Behrens, Ville Kyrki, and Matej Hoffmann. Active visuo-haptic object shape completion. *IEEE Robotics and Automation Letters*, 7(2):5254–5261, 2022.

[15] Edward Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero Soriano, and Michal Drozdzal. Active 3d shape reconstruction from vision and touch. *Advances in Neural Information Processing Systems*, 34:16064–16078, 2021.

[16] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018.

[17] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.